DOCUMENT RESUME

ED 121 862 TM 005 299

AUTHOR	Glaser, Ezra
TITLE	Applying Outcome Measurements: A Guide to Educational Outcome Measurements and Their Uses. Seminar No. 5.
INSTITUTION	Centro Multinacional de Investigacion Educativa, San Jose (Costa Rica).; Georgetown Univ., Washington, D.C. Public Services Lab.
SPONS AGENCY	Agency for International Development (Dept. of State), Washington, D.C.
PUB DATE	Nov 75
NOTE	52p.; For related documents, see TM 005 295-298 and 300
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage
DESCRIPTORS	Budgets; Comparative Analysis; Correlation; Cost Effectiveness; Data Analysis; *Data Collection; *Educational Assessment; *Guides; Information Utilization; Management by Objectives; *Measurement; Measurement Techniques; Program Evaluation; Research Methodology; Research Problems; Seminars; *Statistical Analysis; Test Interpretation
IDENTIFIERS	*Educational Outcomes

ABSTRACT

This guide is essentially designed as a teaching aid for those who would inform planners, officials of educational ministries, school administrators, principals, and teachers about educational outcome measurements. In outline and graphic form, the guide presents topics for discussion in a seminar dealing with the application of outcome measurements. An outline is presented and is intended as a preliminary "lesson plan." A discussion leader, selected by the group, will select materials appropriate for his or her group, adding and illustrating materials that are of particular interest to his or her country's (or region's) educational system. Steps required to use the seminar format are listed. At the end of the seminar, those participating should: (1) Understand the potential usefulness of statistical science to the program administrator, for analysis of program outcomes; (2) Be aware of the problems that statisticians can deal with, and skills that they make available; (3) Understand what the administrator must do in order to enhance the services of statisticians; (4) Be prepared to engage the services of statisticians in the futherance of their own programs and objectives; (5) Be better prepared to critically review reports which rely on statistical analysis and interpretation; (6) Understand the potential of augmenting their own technical skills in statistical methodology; and (7) Be sensitive to difficulties inherent in the collection, analysis, and interpretation of quantitative data, and be aware of the rationale of the technical methods of dealing with them. (RC)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED121862

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Education Outcome Measurement and Their

TM005 299

A GUIDE TO

EDUCATIONAL OUTCOME MEASUREMENT AND THEIR USES

SEMINAR NO. I
Types of Outcome Measurements

SEMINAR NO. II
Uses of Outcome Measurement

SEMINAR NO. III
Generating Outcome Measurements: Achievement and Attitudes

SEMINAR NO. IV
Generating Outcome Measurements: Economic and Societal

SEMINAR NO. V
Applying Outcome Measurements

SEMINAR NO. VI
Feedback Consequences and Steps toward Implementation

Selma J. Mushkin, *Director*
Bradley B. Billings, *Research Associate*

SEMINAR NO. V

APPLYING OUTCOME MEASUREMENTS

EZRA GLASER

PUBLIC SERVICES LABORATORY
Georgetown University
CEMIE (Centro Multinacional de Investigación Educativa)
Costa Rica
November, 1975

This series was prepared as part of a project sponsored and funded under a contract with the Agency for International Development. The views expressed are those of the authors and not necessarily those of the Agency.

PREFACE

This guide essentially is designed as a teaching aid for those who would inform planners, officials of educational ministries, school administrators, principals and teachers about educational outcome measurements.

In recent years, educational services and facilities have made increasing claims on national economic systems. Since the needs of all public services, including education, are pressing and sum to totals far in excess of resources available, it has become urgent that governments initiate processes of questioning both existing resource uses and those proposed, asking: Can we serve the public better and more cheaply? What can we do that would more surely achieve the results we seek?

Responsible education, planning, and finance officials have indicated great interest in measurement of educational results as a beginning step to finding answers to these questions.

In outline and graphic form, this guide presents topics for discussion in seminars dealing with the following subjects:

I. Types of outcome measurements
II. Uses of outcome measurement
III. Generating outcome measurements (achievements and attitudes)
IV. Generating outcome measurements (economic and societal)
V. Applying outcome measurements
VI. Feedback consequences and steps toward implementation.

The outlines for six seminars which follow are intended as guides or preliminary "lesson plans." The discussion leader will select material appropriate for his group. His additions of information and illustra-

tive materials that are of particular importance to his country's (or region's) educational system would greatly enhance the discussion.

"Figures" are presented before each discussion question and are tended to serve as visual aids. The discussion leader may distribute copies to group members or he may find it helpful to enlarge the figures into charts to use for the discussion, or to have view graphs made for this purpose. To facilitate such use all figures are reproduced in appendix B.

The topics covered can be so selected that a general orientation about educational outcomes may be presented in summary form in a single seminar, or more in depth discussion would require at least 6 seminars. Extensive training of one year or more is required to develop skills needed to properly design measurement instruments or the analytical capacity to apply them. However, for those who have the required analytical and quantitative skills already, participation in the seminars outlined would help provide an understanding of the policy implications of the emphasis on outcomes. The seminars are not designed as technical training, rather they are intended to:

- Provide an understanding of the different types of educational outcomes.
- Identify the variety of uses of outcome measurements.
- Help convey the choices on outcome measurement instruments and yardsticks for policy officials, school administrators and teachers.
- Provide some rudimentary understanding of the tools of analyzing the correlates of education outcomes.
- Make plain the possible impact of emphasis on educational outcomes for educational planning and school finance, as well as work in the classroom.

To use the seminar format the following steps are required:

STEP 1
A seminar discussion leader must be selected.

STEP 2
The discussion leader, together with the educational officials sponsoring these sessions should identify the particular purposes of the seminar, and should answer such questions as: For whom are the seminars intended? What is the purpose of carrying out the seminar?

STEP 3
Seminars should be announced as an offering with (a) the clear indication of top level support, and (b) a practical use of the learning

achieved in the continued work of the official, administrator, or teacher.

STEP 4

The discussion leader might select from the materials and "figures" presented in this guide those that appear particularly useful for the specific groups with whom the discussions are scheduled.

As an aid in this selection process, Appendix A suggests some portions of the text that appear most important for the following groups:

- top officials concerned with overall policy planning
- middle management in Ministries of Education and Finance
- school administrators
- school principals
- teachers
- instructors in teacher training institutions.

STEP 5

Supplementary reading materials should be selected. Supplementary readings are available from the Public Services Laboratory of Georgetown University; these include **Educational Outcome Measurement in Developing Countries: An Annotated Bibliography;** and **Educational Outcome Measurement in Developing Countries.**

STEP 6

Some preliminary evaluation should be introduced at the close of the seminars to determine whether the objectives of each seminar have been fulfilled. This can be done by testing instruments, by questionnaire, or by final closing comments of participants.

Such evaluations tell us little about whether the seminars made a difference. Only practical application indicate whether the seminars were successful in changing traditional practices.

TOPIC TO BE COVERED:
APPLYING OUTCOME MEASURES

Objectives of Seminar No. V

At the end of the seminar, those participating should:

(1) Understand and appreciate the potential usefulness of statistical science to the program administrator, for analysis of program outcomes.

(2) Be aware of the classes of problems that statisticians can deal with, and skills that they make available.

(3) Understand what the administrator must do in order to enhance use of professional statisticians.

(4) Be prepared to engage the services of technical statisticians in the furtherance of their own programs and their own objectives.

(5) Be better prepared to review critically reports which rely upon the analysis and interpretation of statistical data.

(6) Understand and appreciate the potential advantages of augmenting their own technical skills in statistical methodology.

(7) Be sensitive to difficulties inherent in the collection, analysis, and interpretation of quantitative data, and be aware of the rationale of the technical methods of dealing with them.

Introductory Questions for Discussion

Several questions are offered, in order to provide flexibility for the discussion leader; the amount of time allocated to this preliminary discussion depends in part upon the amount of time available for the entire seminar.

(1) How might one characterize, in very few summary measures, the target population served by the programs of seminar participants (people, objects, events)? Those actually receiving services from the programs? The program elements themselves?

(2) How might one summarize the relationships between the program activity (i.e. the actual delivery of services) and the changes which actually occur (or differences which are created and maintained)?

(3) How might one separate the results of one educational program from other factors and influences that might be operating at the same time?

(4) How might one attribute an observed change (say, a desired program outcome) to (a) specific operations of programs, (b) interfering factors, (c) variations or changes to which no specific cause or influence can be assigned?

(5) What are the most pervasive problems in attempting to measure the program outcomes of the participant's programs, and to relate these outcomes to the activities and characteristics of the programs themselves?

NOTES FOR THE DISCUSSION LEADER

The basic theme for the seminar is that there are formidable difficulties in the collection, analysis, and interpretation of data. It is this natural inherent unavoidable difficulty that creates a need for formal statistical methodology. Moreover, it is in precisely program accountability and output programming in its various forms (PPBS, Management by Objectives, etc.), that require such data interpretation.

A correlary theme of the seminar is that administrators can draw on professions and combine help from two professional sources. The first is represented by professional competence in the substance of a particular program: such as education. There is no way to approach the analysis of programs without a firm base in the theoretical and applied foundation in the relevant subject matter. The second professional source is statistical science: the formal mathematics-based methodology aid in measurement problems, such as social surveys, and measurement.

A third and final theme is derived from these two: that it is not a simple matter for the administrator of a public program to obtain these technical

resources and turn them to his own advantage. *It is precisely this need that is met in this seminar:* to better prepare the administrator to conduct his affairs with the well focused assistance of professionals.

There are seven parts in the next section. Each frames the generic problem that is frequently met by the administrator, and presents the germ of the statistical methodology that is available to assist in its solution. The seminar can be enriched by further examples from the experience of participants. Specific application to the programs of seminar participants presents the principal learning challenge and would serve to reinforce the essential points.

DEVELOPING AND USING OUTCOME MEASUREMENTS FOR PUBLIC PROGRAMS

Generic Problem	*Analytical Methods*	*Typical Applications*
Characterizing and comparing groups of people, objects, events.	Totals, subtotals, means, distributions, tests of differences between means, etc.	Detect changes in the character or results of programs; differences over time, between geographic areas, between program technologies.
Measurement of relationships.	Contingency tables, scatter diagrams, correlation and regression analysis.	Discovering factors that influence the outcomes of programs for better or worse; selection of program elements for superior results.
Isolation of influences; decomposition of trends or output measurements.	Trend analysis; regression analysis; "control" by experimental design of unwanted interferences.	Attribution of program effects and consequences in the presence of intervening or interfering factors.
Measurement of variation and accounting for its components.	Variance analysis; measurement of the components attributable to program factors; to non-program (environmental) factors; to interactions among factors, and to "experimental errors".	Measuring the relative importance of program factors, environmental factors, their interactions, all of these components interpreted against explicit measure of the sensitivity of the measurement system.
Collection of Data to precise specification of an analytical model and analysis plan.	Survey methods, instrumentation, adaptation of administrative procedures and records, etc.	Gathering data about target populations, program operations, costs, direct (immediate) outcomes, and (possibly remote) program consequences.
Obtaining required information economically and to high quality standards, with self-contained measures of precision.	Sampling and experimental design, with quality control of field operations.	Reducing costs; reducing response burden and interference with operations; maintaining high quality.
Measuring program outcomes and consequences, assigning them values, choosing program elements and technology with highest output per unit of resource inputs.	All of the above techniques plus evaluation schemes for converting outcomes and consequences into social benefits, including planned variation experiments.	Measure levels of outcomes; changes over time; relation of outcomes to program factors (cost-effectiveness), and comparisons of different programs by outcomes (benefit and benefit-cost analysis). Many applications to planning, budgeting policy analysis, etc.

Ideas for Presentation

This section is organized around seven generic problems of acquiring, analysing, and interpreting data for management uses. The sequence is deliberate. In the actual operation of a reporting system and the use of structured data, the procedure is: (1) make inquiries and collect data (application forms, survey questionnaires, administrative records, tests (written and oral) or various forms, recording of instrument readings, etc. (2) "process" the data in a manner that reflects the form of the analysis to be performed; (3) proceed with the formal analysis; (4) interpret the results of the analysis—elaborated below—and apply the conclusions to the administrative problems, plans, and actions of the particular organization and program.

It is important to note that the planning of projects proceeds in the opposite direction. First, there must be specific needs based upon priorities of the administrator and his superiors. These needs dictate the content and form of the statistical analysis and the manner in which results must be interpreted in order to be useful. The measurements, relationships, and sensitivities essential to these formal analyses dictate the manner, extent, and content of the data acquisition process *in full detail;* that is, what is to be collected, in what manner, whether by sample (and what size and content of sample), etc.

There is another way of proceeding. Information is acquired without first carrying out the steps which should have been planned in detail in advance. Often a "data bank" is built up by these means; the plans for analysing, interpreting, and using the information are not explicitly taken up until the data bank is established. We say "explicitly" because there are many ideas about use and analysis inherent in the choices of the information to be collected, in what form (including classification structures for subtotals, etc.), and from what sources. However, experience has shown that the administrator's need is too precise and too demanding to start with hunches about information, in the unspoken hope that the necessary relationships, etc. will all be somehow derivable from such a rich "data bank". Perhaps the poor record of analysis and appraisal of public programs can be more frequently accounted for by this failure of strategy than from any other non-policy factor.

Problem No. 1: Data to characterize a group of people, objects, or events, and to make valid comparisons among such groups.

Data are frequently required to provide a summary description of a group such as children of school age in a particular area. Often this summary is merely the recorded or estimated number of individuals in the group: the total. There are probably instances where this summary will suffice, for some limited purpose, so no further information is necessary. More frequently, further meaningful questions arise. How many of those school-age children were attending school? Of those who were not, what is the highest grade attained? These answers are only comprehensible if related to the specific ages of the children, and it is usually useful to have separate answers for boys and girls.

The ability to divide a group into subtotals implies that there is a well-defined *classification* system, with unambiguous *definitions* of the classes.

We will also see below that many of these answers can be interpreted usefully only if it is possible to *relate* one subclass characteristic with another; the subtotals for ages of children and school attainments leave too many important questions unanswered. It is necessary to cross-classify age and class so that one might *distribute* each level of attainment by age (and probably sex), and each age by school attendance (present grade) and attainment (for non-attenders).

To summarize, we can characterize a group by presenting the grand total number of individuals, and by classifying them for presentation of subtotals, which means that we can *distribute* them into subclasses by means of a classification system. We can also relate any subtotal with any other to form a variety of distributions according to the attributes of the members of the group.

It is also possible to distribute a group according to a *variable* or a scale of values, such as the height of children or the test scores for a school reading test. For both practical and theoretical reasons, a useful summary of a batch of test scores or heights is the *average* or arithmetic mean. Hence this single summary statistic can be presented to represent something important about the group. As in the case of a total, however, the single figure leaves a good deal

unaccounted for.

A first refinement is to present the averages for each of the sub-groups found to be useful in the component analysis by subtotals: separate test scores (or heights) for boys and girls separately for each year of age, for each present grade for those attending school, and for highest grade attained for non-attendees.

It is also possible to develop a distributional analysis for a single subclass of individuals: say girls 10 years of age and presently attending the fourth grade. They have an average score on a particular test, but not all of them got the average score. (Indeed, it is possible that none of them did so.) We can distribute the scores of the individual girls in this subclass by merely counting the number that received each grade, and showing the result graphically. (Figure V-1). It is also possible to indicate the fraction of scores that fall within certain distances from the average: say, within five points of the average, ten points, etc. providing a new array of measurements or statistics which augment the average as a statistical description of the group. Technical statistics depends heavily upon such measures of dispersion about class means, because the mathematical theory of probability provides some highly useful mechanisms for interpreting them. (The statistician will use the measure of "variance", based upon the squared differences of each score from the mean.)

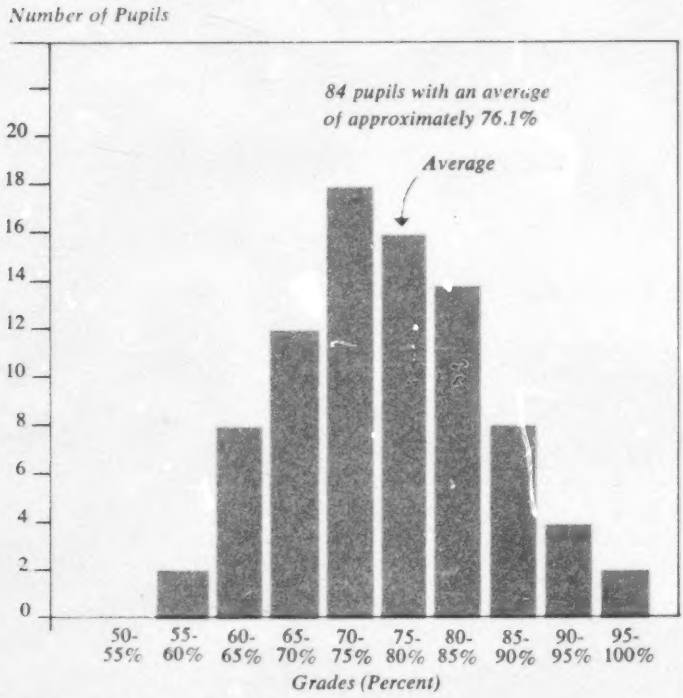
We now have some of the essential machinery for analytical comparison of two or more groups. Returning to the 10-year-old girls, it is now possible to compare present 10-year-olds with a group that was 10 years old some specific date in the past, in order to determine whether school attendance and school attainment are improving. It is also possible to compare one community with another. If standard achievement tests are in use, the same comparisons can be made for 10-year-old pupils in different years or communities.

The use of totals, subtotals (based upon a well-defined classification system), distributions of attributes, and distributions of variables can often make possible all of the useful comparisons that are essential.

The administrator of an educational program will frequently have an additional problem: analysing differences which he would like to attribute to his own program. To illustrate, the introduction of a new textbook is expected to improve the test scores of the children. The need is to compare those who had this advantage with other communities and earlier classes who did not. For sensitivity, it is

FIGURE V-1

TEST SCORES OF GIRLS OF AGE 10 YEARS WHO ARE IN FOURTH GRADE (Illustrating a Distribution)



often necessary to subgroup the children in various ways (age, sex, etc.) in order to understand the impact of the improvement.

We shall return to this problem below but a useful start can be made with the relatively simple statistical machinery sketched out above. It will be seen that distribution theory plays important and subtle roles in the interpretation of experimental data, survey data, and sample data; it will usually be necessary to rely upon the training and skills of a technical mathematician for a valid interpretation of the analyses.

Problem No. 2: To measure the relationship between two factors or characteristics; in particular, to measure the possible outcomes of a program, and to associate it with some measure of delivery of that program.

We begin with an elementary example (which will turn out to require some refinement in most real-life instances, as will appear in Problem 3). Suppose that we experimentally introduce a new program in which third-grade pupils are given free textbooks in certain schools but not all schools. We should like to measure the influence of the textbooks on the retention of students for the next grade. We will take all of the pupils in the third grade with the free textbooks, and about the same number of pupils in the third grade from other schools. Our measure of program effectiveness will be the percentage of this year's pupils who matriculate in the fourth grade next year.

The results can be summarized (hypothetical data) in a simple contingency table:

Schools	*With Free Textbooks*	*Without*
Percentage of third-grade pupils who matriculated in the fourth grade the following year	80	61
Percentage who did not	20	39

The table would seem to demonstrate that 19 percent more of the original third-grade pupils were retained as a consequence of the free textbooks. (Questions might arise about the presence or absence of other differences between the schools in the program and the other schools for which data are compared: this possibility will be explored in Problem No. 3. One might also ask how the experimental schools were chosen: why these particular schools? This technical issue will be treated in Problem No. 6.)

In an example from medical science, we are told that all of the sheep Pasteur vaccinated for anthrax survived in good health when

mixed with diseased sheep, but that the unvaccinated sheep all died. The result:

	Vaccinated	*Not Vaccinated*
Lived	100%	0%
Died	0%	100%

No one could argue with such a result. But modern medicine (and education) often grapple with difficult problems whose mitigation is not so simple. Suppose Pasteur had tackled one of the more difficult medical conditions and it produced the following results:

	Treated	*Not Treated*
Lived	41%	37%
Died	59%	63%

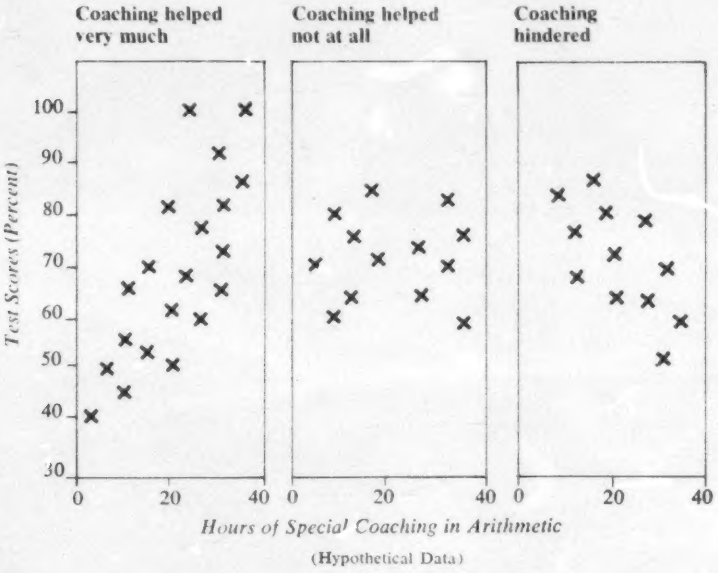
What is the possibility that the four-percentage-point difference could arise, in favor of the treatment, where the treatment really had not any effect at all? This question can be answered by the technical statistician by the use of the mathematics of probability. It will probably turn out that the difference is a real one—even though the treatment is not very effective—only if there are very large samples—and random samples—of treated and untreated cases of the disease. Note the same basic logic as in Problem No. 1, where the requirement is to measure the difference between two groups of individuals by the use of means and distributions.

If we move from attributes (which can either be present or absent, like vaccinations and survivals) to variables (which can take on a number of values), some more powerful techniques become available.

Suppose we have a special coaching program in arithmetic, for a particular grade, with the pupils being coached by individual tutors for varying numbers of hours, with some not coached at all (i.e. zero hours). Suppose that a test is then given and the scores recorded in such a way as to permit each score to be associated with the number of hours of coaching the pupil received. We can then construct a *scatter diagram* where each point corresponds to one particular pupil whose test score is measured by the scale on the left (vertical distance) and whose number of hours of coaching is measured along the bottom scale (horizontal distance). The left panel of Figure V-2 would correspond with the situation where the coaching helped the pupils very much; the center one very little

FIGURE V-2

RELATION OF ARITHMETIC SCORES TO HOURS OF COACHING (Examples of Scatter Diagrams)



(Hypothetical Data)

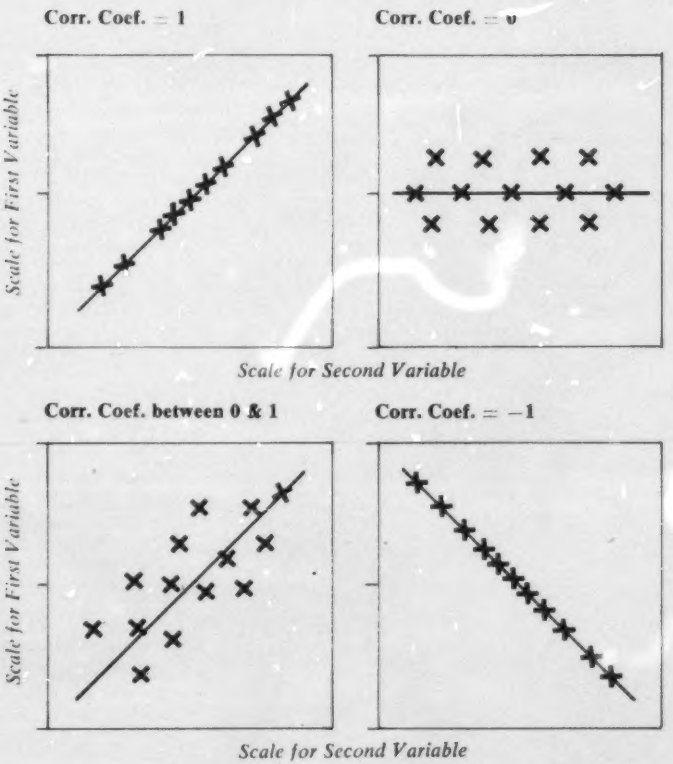
or not at all; the right panel shows the situation where the coaching actually impeded the ability of the pupil to score well on the test. (All hypothetical data).

How might these relationships be summarized with suitable statistics? There are two closely related forms of analyzing relations of this sort. The first is called *correlation analysis;* it measures the correlation between the two variables. In Figure V-3, the left panel shows a correlation that would be calculated as a *correlation coefficient* of 1, the second panel would compute out as zero, the third would be somewhere between zero and one, and the last would compute out as minus one. (The coefficient is computed as the ratio of squares of differences from line drawn in each panel to the squares of the differences from the arithmetic mean; the line is the

FIGURE V-3

CORRELATION BETWEEN TWO VARIABLES

(Illustrating the Correlation Coefficient)


Corr. Coef. = 1
Corr. Coef. = 0
Scale for First Variable
Scale for Second Variable
Corr. Coef. between 0 & 1
Corr. Coef. = −1
Scale for First Variable
Scale for Second Variable

(Hypothetical Data)

"least squares" fit because any other line drawn on the same scatter diagram would have a larger total of squared differences than the one drawn.)

The second form of analysis is called *regression analysis*. It utilizes the same "least squares" line—the line of "regression"—to estimate what the test scores would have been if the group had been given any number of hours of coaching. Note that a single pupil will yield this estimate only if there is a perfect correlation of 1.0. For other cases, there might well be not a single pupil with the specified number of coaching hours who earned the score now estimated to be "most likely" for that number of hours.

It is possible to study more than two variables simultaneously with these techniques. The "multiple correlation coefficient" might express the influence of number of hours of coaching, class size for the normal classes, and years of experience of the arithmetic teacher. The regression analysis would permit the estimation of the test score for any combination of hours of coaching, class size, and teacher experience.

The application to program analysis is perfectly straightforward. One must select a suitable population of pupils (Problem No. 6), devise a scale which might be used as a measure of success (Problem No. 5), and arrange for variation in the factors to be studied (No. 6 again) so that the influence of each one can be measured against the chosen measure of success (outcome). There is a special situation where two factors are influential, but only if both of them are present. Regression analysis cannot measure this necessary-but-not-sufficient condition for single variables; this kind of "interaction" requires a different statistical technique (Problem No. 4).

Problem No. 3: To separate or isolate the influence of a program from interfering influences that might cause similar results.

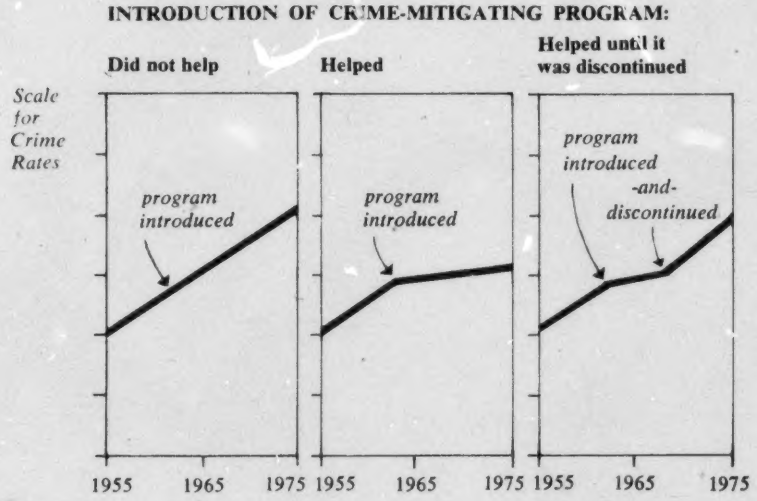
This is a frequently met situation in the real experience of program evaluation thru the analysis of outcomes. The world simply does not hold its breath when some public authority decides to attempt to mitigate or remedy a social problem. The changing world continues to change, and trends often continue, even if modified by the new public program.

How, then, to isolate the effect of the program from other changes?

FIGURE V-4

EFFECTS OF CRIME-PREVENTION PROGRAMS

(Illustrating Trend Analysis)



(Hypothetical Data)

One approach is thru *trend analysis*, a study of the form of changes over time. The nontechnical press frequently carries such comments as "after spending millions on crime prevention, the crime rate continues to increase: therefore it is obvious that the program is a failure and the taxpayer's money has been wasted". The statement would perhaps be justified—in an elementary and far-from-complete analysis—if the left panel of Figure V-4 were the case. But in the second panel one might question that the money was "wasted", notwithstanding that the crime rate continued to increase. What this graph says is that there are such powerful influences in the direction of increased crime (changing demographic characteristics, economic conditions, and social attitudes might be worth study, as well as public safety, criminal corrections and rehabilitation and related programs), that it continued to increase in spite of a short fall in the *rate of increase* coincident with the institution of the new program. Perhaps—in a more complete analysis—the third panel would be even more persuasive. What it says is that the program mitigated the rate of increase, but without lasting effect: the removal of the program coincided with a resumption of the earlier trend, with an augmented rate of increase in the crime rate.

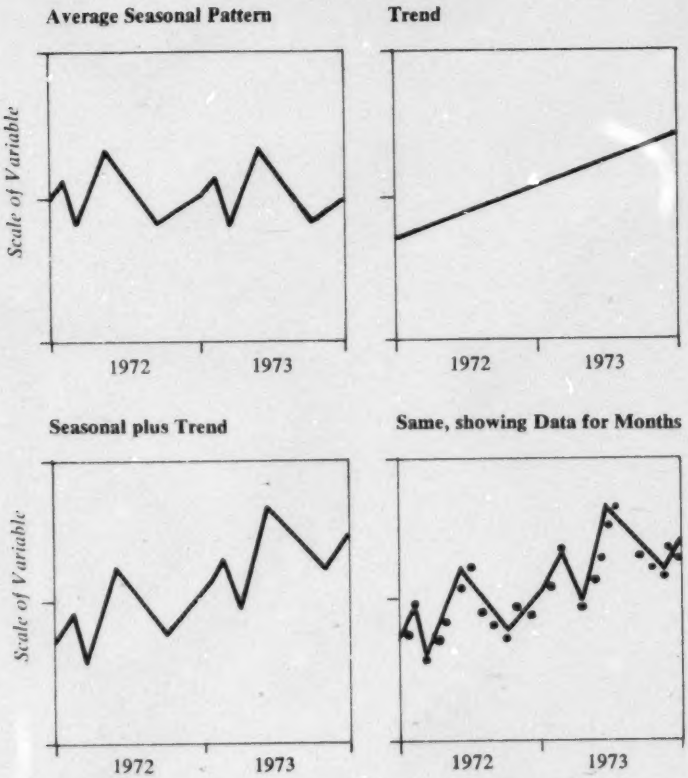
Education has its share of loose and inconclusive conclusions, masquerading as the results of serious analysis: "After spending all of the money on Title I (of the U.S. Elementary and Secondary Education Act of 1965), schoolchildren still cannot read." What is needed here is (1) more precise definitions and quantification of the notion that "children cannot read"; (2) measurement of the changes in this measure of reading ability; and (3) isolation of the effects of the Title I program (with its many and varied elements) from other coinciding influences.

Trend analysis, then, relies upon coincidence in time for its ability to attribute "causes" to program factors and other influences. A more advanced approach involves the separation of components of a time series into its components. For example, the first panel in Figure V-5 shows a regular pattern of change thru the average year —a "seasonal" component; the second panel shows a trend (possibly computed over a period of several years; the third shows the combination of the two components; the fourth shows an actual time series (hypothetical data); and the final panel shows the *residual variation* or unexplained variation after taking account of an average trend and an average season pattern. It must be attributable to: (1) the inability of the *average* trend and the *average* seasonal

FIGURE V-5

VARIATIONS OVER TIME

(Illustrating Components of a Time Series)



(Hypothetical Data)

pattern to describe what happens in a particular year—i.e. the *distribution* of data around their averages (c.f. Problem No. 1); and (2) all other causes and influences other than trend and seasonal patterns. The study of "business cycles" in economics has made refined use of such techniques, finally demonstrating that the regularities of timing were too unreliable for satisfactory utilization of this strategy for business and economic forecasting.

Other strategies are available for isolation of unwanted factors in the study of program outcomes and for the evaluation of public programs. Data may be collected in such a way (Problem No. 5 below) that one can isolate one factor from another. This might be accomplished by constructing a contingency table (Problem No. 2) and permitting its needs to dictate the specifications for data acquisition (Problem No. 5 again). Suppose that a particular community has some schools that typically score high on standard tests of achievement, and others that score low. Suppose that it is known that the school with better scores also has most of its children from relatively high-income families with well-educated parents, and that the low-scoring schools have pupils that more typically come from lower-income families with parents of less educational attainment.

Is it the family background or the schools that cause the difference in the performance of the children? If one merely selected a group of children for each class of schools (even if this were done in some "fair" and impartial manner, as will be discussed in Problem No. 6), the results would most likely take on the appearance of the contingency table in the first panel of Figure V-6. The whole interpretation would rely upon the few children that came from low income families but attended the better-scoring schools and the few who came from wealthier families but attended the poorer schools. The rest of the cases would have little information, since the factors that were to be isolated had been "confounded" in the design of the data-collection. A technique is available to increase the size of the sample in the cells that discriminate between the two kinds of influence, and reduce the number in the cells that confound the two. The second panel shows the result in terms of sample structure. The final panel shows the average test scores for each cell, with the highest reliability in the two cells that discriminate between the family backgrounds and the "quality of the schools". In this hypothetical example, the analysis would show that it was the family background rather than the

FIGURE V-6

TEST SCORES OF CHILDREN FROM SCHOOLS IN TWO COMMUNITIES

(Illustrating "Confounding" of Explanatory Influences and Sensitivity of Statistical Analysis)

Children whose Parental Educational Attainment is:	*Schools from Communities where Average Family Income is:* **High**	**Low**
High	214: 83%	2: *
Low	4: *	307: 72%
High	50: 83%	200: 81%
Low	200: 75%	50: 72%

Key: Each cell shows number of children taking the test and their average score

(Hypothetical Data)

**Sample too small for average to be significant*

characteristics of the schools that was associated with the differential achievement. This could have great importance in policy making, budgeting, and the rational administration of educational programs.

We have avoided two problems in the above account. First, how are we to know about the family backgrounds before collecting the data? This problem can be solved by designing the data collection in two stages. The first stage determines the family backgrounds from such a large number of children that the final sample in the second stage can contain as a majority of the cases the kinds of children that make up only a small fraction of the school-children. The second is that we combined two variables in our description of "family background": income, and educational attainment of parents. Again, this kind of confounding can be resolved by a preliminary stage of data collection. Although the majority of the parents in higher-income families would almost certainly have more education than those of the lower-income families, there must be many combinations of high-income/low-attainment and low-income/high-attainment to build up the sample in contingency tables that separate these influences.

There are other strategies for isolation of influences that might confound or otherwise interfere with measurement of program outcomes. These involve the "control" of possibly interfering factors by specifically designing the analysis and data collection to permit the *analysis* to isolate factors, even though one cannot find *individuals* that exemplify discriminating cases as in the contingency tables above. This will be discussed in Problem No. 4 below.

It should be realized, that this common-sense approach to statistical analysis rests upon a firm theory of mathematical statistics, which can protect the analyst from being deceived by the data in many ways. What has been presented here is a basic logic of the methods. In practice, the analysis can be full of traps for the untrained and unwary (remember Problem No. 1). The intention here is to sufficiently train the seminar participant so that he is aware of the strategies that the trained statistician can bring to the solution of the administrator's problems, and—most importantly—how to instruct the statistician concerning the needs of the analysis to serve the useful purposes of program analysis, budgeting, and related operations.

Problem No. 4: To measure the total variation in some scale of values (say, an outcome measure), and to assign the components of that variation to program influences and to other associated factors. Statistics is the systematic study of variability. If there is no variability, there is nothing to study.

Consider the problem of schoolchildren performing poorly in all school situations. If the unimaginable condition described here actually occurred, analysis by quantitative methods would be impossible: all children received the same grades, test scores, and other evaluations all of the time. Since there was no variation, there was no way to assign any influences to success in education.

In the real world, one child differs from another in a great variety of ways. Our problem is to impose a formal structure on these variations in such a way as to permit us to understand as much as possible the factors that account for the variations.

Let us suppose that a reading test has been given to three schools, for all of the children in the third grade. Let us further suppose that we have no particular information about the schools, but they might differ in a number of ways, each being in a different community. The children also differ in many ways, but all we know about them —permitting these factors to be associated with the individual scores—is their *sex*, *age*, and the *educational attainment of their father* (or their mother, if the father does not live at home).

We would like to construct a framework for analysis—a *"model"*—which sets forth the measurements and relationships that will permit the "explanation"—the *association*—of differences in scores with the various factors noted above:

- Differences associated with differences between the schools (even though there is no record concerning what these differences might be);
- Differences associated with whether the children in the third grade are girls or boys;
- Differences associated with age of the third-grade pupil;
- Differences associated with differences in the educational attainment of the parents.

Notice that only the first will probably be considered a "program" factor, in that it can be manipulated by changing budgets, by

training or retraining teachers, by giving out free textbooks or by providing new ones, etc. Measurement of the differences between schools is only the first step in the process of attempting to learn what causes the differences and using the information to improve *all* of the schools to the extent that educational technology (books, other educational materials, teaching methods, etc.) and resources permit.

The associations of differences in test scores with differences in sex, age of the pupil and educational attainment of parents are measured, not because the administrator has the ability to manipulate them, but because it is essential to avoid confusing the various influences. It is also instructive to actually measure the fraction of total variation that is under the control of the educational system, rather than part of the environment in which the system must operate.

The above list of sources of variation is not complete. It lacks measures of *interaction*, and the *residual variation* (measured as residual "variance"), which can usefully be regarded as "experimental error" in that it indicates the limits to the ability of the model to explain what is happening.

Interaction differs from such measures as difference-between-schools. In the latter, the analysis measures the *isolated effect* of one of the factors in the model (a particularly important one in this example). The interaction measures the effect of two or more factors working together. In an oversimplified model, we might have arranged to have classrooms without any teacher, and teachers without classrooms, blackboards or teaching materials. Study of test scores under these conditions almost certainly would show little advantage in having educational materials without teachers, and only limited success for teachers without teaching materials. But the *combination*—the interaction of teachers, schoolrooms, blackboards, and textbooks—would be expected to be associated with higher test scores.

Another picture of the meaning of interaction can be obtained from the following simple illustration:

Average Test Scores in Arithmetic

	Ethnic Group A	*Ethnic Group B*	*Both Ethnic Groups*
Boys	40	60	50
Girls	60	40	50
Both Sexes	50	50	50

The grand arithmetic mean for both sexes and both ethnic groups is 50 percent. Each sex and each ethnic group also has an average score of 50 percent. But the pattern of scores indicates that these statistics fail to demonstrate that something more is occurring.

There is something about being a boy in Group A or a girl in Group B (in this hypothetical example) which leads to scores much higher, on the average, than the other two combinations. These statistics merely describe the *association*, not the *causes* of the differences. However, there is no reason to search for causes until it is clear that there must be at least one. (A conjecture to be tested might be that the ethnic groups attended different schools, and that one school had a teacher that particularly appealed to boys and the other had a teacher that particularly appealed to girls, with the attitudinal and motivational factors of this situation finding expression in the pattern of test scores.)

Our strategy will be to measure the *total variation*, and then to subtract the variation associated with each of the isolated factors, and the interactions between each combination of two factors ("first order" interactions), and the interactions among more-than-two factors at a time ("higher order"). After subtraction of these components, there will always be some of the original variation left —unassociated with any of the above increments. This is called "residual variation". It represents all of the influences of factors not included in the model, and all of the interactions not included. These must encompass all of the differences in the ability of the children, their learning skills, attitudes, motivations, attendance records, household influences not related to the educational attainment of the parents, physical condition of the child, etc. It also includes all variation associated with the test itself and with the testing procedure.

The residual variation serves two vital purposes: (1) it reveals the relative importance of the factors whose influence has been measured, and all other influences on the test scores; and (2) it provides a base for measuring the sensitivity of whatever differences are actually measured. In the latter purpose, we might doubt the validity of a 2 percent of the variation associated with differences between schools, if 65 percent of the total variation could not be associated with any factor or interaction in the experiment. By contrast, a 15 percent difference between schools, with a residual variation of only 10 percent would seem to be important. (The technical statistician will fall back on the mathematical theory of

probability for a more exact interpretation of data of this sort, and the sizes of the samples will be influential in the interpretation.)

How are we to scale the total variability, and its various portions, as described above. There are important theoretical reasons for using the "variance" from the arithmetic mean as the scale: obtained by taking each individual score (say 73 percent) and taking the difference from the mean (say 3 percentage points if the mean is 70), and squaring that difference (9 in this case); and adding up all of these squared differences. Similar computations will reveal the differences between school averages, ages, etc.

There will always be a residual "variance", the remaining portion that the experiment does not explain.

There are many potential applications of "*variance analysis*"– the technique set forth above in the administration of public programs of education and for other social services and for the measurement of outcomes. More importantly, it becomes possible to associate measures of outcomes with characteristics of the program and its operation—especially with some of the resources or inputs to the program: funds, staff, physical facilities, materials, etc. (The implications of this ability will be further discussed in Problem No. 7.)

Problem No. 5: To determine what principles govern the collection of data for all of the types of analysis described above.

We are now in a position to discuss the problems of data acquisition, having reviewed a number of types of analysis that will prove useful for the measurement of outcomes of public-service programs and educational programs in particular.

The first requirement is knowledge of the subject to be investigated. A highly skilled chemical engineer will probably be unaware of the important characteristics of the elementary education system, and the needs for program assessment—in the detail required for disciplined statistical analysis. Even the specification of a "third-grade arithmetic test" or "educational attainment of the parents" require careful conceptualization and precise definition. These matters are not the subject matter of theoretical statisticians (although those who work in applied statistics are well aware of the pitfalls, if not the answers); this is the training and experience of professional educators and educational administrators.

What arithmetic test is to be used? What is known about its ability to discriminate among pupils near the middle of the distribution? Is it in use in the areas with which comparisons might profitably be used? What are its known biases? With what population has it been "normed"?

Is educational attainment to be described in three classifications (no schooling, completed three grades of elementary school, completed at least four, e.g.) or ten? Does the conceptualization extend to on-the-job training? Self study? Nonformal methods, such as the use of television? Are any substitutes to be used, such as a job which requires a good deal of reading and calculating skill?

After listing the pieces of information that will be needed to satisfy the demands of the analytical model, defining and formatting the items, preparing survey schedules or interviewer worksheets, etc., there is yet much to be done. For interviewing, field staff must be recruited and trained. For mailed questionnaires, there must be proper covering letters, with persuasive reasons for the respondent's cooperation, and a complete follow-up routine for the initial nonrespondents. For obtaining data from school records and other

administrative files, the clerk must be trained. All of this procedure must be carefully supervised and various quality maintenance measures must be utilized (such as a formal "quality control") to be sure that all instruction are being faithfully carried out.

It is by no means rare that an otherwise competent research project fails for lack of care in this part of the work. Expert survey organizations not only give a good deal of attention to all aspects of the data collection; they also arrange to *prove* that the field staff followed the prescribed procedures, that the respondents understood the questions, that they had adequate basis for the information they provided, etc. The post enumerative quality check has become a standard part of properly conducted census surveys, for example. A small highly trained staff reinterviews a randomly chosen small sample of actual respondents with inquiries in depth to establish the validity and accuracy of the data.

Even information from administrative records require quality assessment, for there are errors here also, and also persons making entries who do not fully follow the official instructions regarding these records.

Problem No. 6: To arrange for the collection of statistical information (i.e. structured quantitative information) in such a manner that the necessary information will be acquired economically, that it can be used for generalization, and that its precision can be assessed.

Sometimes there is interest in a particular group of children in a school. There is no issue about the selection of cases if the number in the group is sufficiently small to include all of them without undue cost or administrative problems. The resulting information, analyses, and interpretations will pertain exclusively to that particular group of children and no others.

Much more frequently in the administration of public service programs such as education, the inherent interest is in very large groups—too large to observe all of the individuals. Sometimes the interest is in groups that cannot be observed, such as future third grade pupils, or all present and future third graders utilizing a new array of educational methods and materials which have not yet been routinely introduced. The problem, then is to draw conclusions about individuals from which no data were collected.

(In activities other than education, there are clear instances of "destructive testing" in which the individuals observed are destroyed or are "used up". This occurs in educational testing in a more subtle form: those who took a particular examination cannot be used to represent children who have not had this experience, if it is judged that the testing procedure involves learning or practice. A clear case is in seed planting. If all were tested in order to obtain the germination ratio, there would be no seeds to plant for commercial crops. It is therefore necessary to devise a way to plant a relatively few seeds and to draw valid inferences about the whole supply of seeds from the behavior of the small collection.)

Very often the costs, time, disruption of school routine, and the need for very careful carrying out of procedures dictate that conclusions concerning a large group will be derived from observing a small part of the group: *a sample*.

But educators already know that children vary widely in all kinds of traits, including some traits that are difficult to conceptualize and to

measure. And what about characteristics that we do not know about that might importantly affect the performance of the child?

The critical principle is to draw the sample *randomly*, so that every individual in the group has an equal chance to be chosen. The choice must depend strictly upon *chance*. No element of judgement may enter. No selection mechanism can be allowed to enter, which may bias the results. People who answer mailed inquiries may be significantly different from those who do not, with this difference affecting items of information sought by the survey. Household visits can be biased simply by having the field staff avoid muddy entrances to houses or houses with unleashed dogs in evidence.

With each member of the *group about whom inferences are to be drawn* equally likely to be chosen, it is possible to draw inferences about the entire group, from the random sample that actually provides information. This applies equally well to administrative records, test performance, information about households and families, and other kinds of information and formats of the information.

Several critically important characteristics of strictly random samples:

Comment No. 1: The sample should be unbiased. No judgement of analysts, experts, or political authorities has influenced the choice of those who will provide information (or about whom information will be compiled). Suppose there is a personal characteristic which has not been identified as important in test performance of school-children: we can merely designate it as "Characteristic A". Suppose, that 38 percent of the children in the group being studied—and sampled—have this characteristic, and the others do not. Then the random sample will also have 38 percent of its individuals with Characteristic A (subject only to sampling variance, discussed immediately below), and therefore be *unbiased with respect to this factor even though no one knew its importance or even that it existed!*

Comment No. 2: The mathematical theory of probability permits the statistician to compute the *precision* of a sample from its size—i.e. the number of individuals sampled— *provided that it has been randomly drawn.* The total number of people in the group being studied—those included in the sample as well as others do not affect the precision of the sample and its statistics, though this is intuitively disturbing to many nontechnical observers. (A rare exception is the case of a small group being studied, and the sample

comprising most of them.) The precision is dependent only upon the size of the random sample and any degree of precision can be attained by making the sample large enough (the "law of large numbers", for random samples only, and not dependent upon the distribution being "normal" or any other particular form). Hence the technical statistician will ask what precision is needed, so that the necessary sample size can be computed *before the rest of the detailed plans are developed*, and well before the actual collection operations.

Comment No. 3: In practice, the sample often has "structure". In Problem No. 2, there was an example of two-stage sampling in order to obtain enough cases of poor children in schools with generally wealthy children and the reverse combination. In such an analysis, the precision of *individual cells* in a contingency table or other tabular analysis can be established in advance and the collection of data planned accordingly. In general, then, the size of the sample will depend upon the smallest subgroup in the population (or people, objects, or events) *about which inferences are to be drawn* and conclusions to be reached on the basis of the data collected.

Comment No. 4: It obviously becomes absolutely essential to develop a firm definition of the group to be studied in advance of designing a sample or a study based upon a sample. (It is equally important to do so if there is to be complete coverage of the group —a 100 percent sample or "certainty" sample; otherwise there will be no way to identify the proper respondents or to control the collection of information.) It is also necessary to develop the "analytical model" with its intended comparisons, controls, etc. so that the sample will have the required "structure", i.e. there will be sufficiently large subsamples to support the intended analysis and interpretations of the data.

Comment No. 5: The ability to protect data from bias, with random samples and proper control of data collection, should not be confused with the isolation and control of factors by making them an explicit part of the analysis and the data collection, as in the case of the analysis of variance attributing influences to factors and their interactions, in Problem No. 4, above. In that instance, we had factors defined in advance, *whose influence we wanted to measure*. In the case discussed immediately above, the intent was merely to prevent the factor (known or unknown) from destroying the validity of the analysis.

Comment No. 6: By now, it should be abundantly clear that the design of a statistical study requires a large measure of specialized skill and knowledge. This is what the trained statistician brings to his work. But it is also essential that the administrator or the customer for a statistical study *know what he needs:* what kinds of groups must be studied, how they must be classified and divided into subgroups, what factors and what relationships will have to be developed in order to apply some of the techniques discussed earlier in this review of statistical technique. He must also have strong ideas about how precise the data must be; for illustration, how small a difference between groups must be detected (possibly the difference between children who have had some new educational service or advantage and those who have not).

Comment No. 7: It should also be evident by now that a high-quality study can often be accomplished by keeping the sample to the smallest size consistent wtih the needs of the study and spending the available resources in the assurance that the sample has been randomly drawn, that the field work has been properly conducted, that the answers or information has a high validity, and that all operations have been carefully controlled. It seems anomalous to some observers that a sampled survey (or other data collection) can be far more valid and accurate than a study that studies all of the individuals about whom conclusions are to be drawn, the burden of collecting so many cases—with a fixed budget—invites all manner of *non-sampling errors* which might have been controlled (and *proved* to have been controlled) with the same budget, and a better designed study.

(This might be a rewarding time for the reader to return to the first section to review the objectives of the seminar. They will probably have important new meanings for him, unless he had the advantage of substantial statistical training in advance of the seminar.)

A final point is that the administrator, in our present context, is the chief, the boss, the authority. He must decide what is important. The technical staff has the responsibility to see that he gets it, if it is possible with the available resources.

The entire process of developing the questions to be answered, the groups to be studied, the quantities and relationships to be measured, the interactions to be measured, the precision to be met, the complete analysis plan to be used, the size and structure of the sample (including the possibility of complete coverage), is called the

"experimental design". In this sense, the "experiment" might be merely giving some children a test of performance in their schoolwork, or it might be merely collecting information about individual children, teachers, and schools. However, it might also involve the *random selection of cases* to be "treated" and "not treated". In an attempt to discover whether a new textbook is better than one already in use, it would be necessary to randomly choose the children—better: randomly choose the *classes* to get the new textbook—so that inferences can be drawn about *all children* in the particular school system. It will also be necessary to protect against foreign influences destroying the comparison, and it would be highly advisable to measure the teacher-textbook interaction in order to determine what kinds of teachers do well (if any of them) with the new book—and what kind of characteristics and training the teachers must have to obtain the full benefit of any improvement that turns out to be possible.

Problem No. 7: To apply the discussion of statistical principles t output-oriented management: to management by objectives, output budgeting, cost-effectiveness studies, and benefit-cost studies. The key strategy is to define and measure outcomes—outcomes of education in the present context—and then to utilize this structured quantitative information in all possible ways for the planning, budgeting and administrative of an efficient, effective, and successful educational enterprise.

Probably the most pervasive innovation in public administration since World War II has been the emphasis on formal techniques of management *based upon the measurement of outputs—outcomes—and the relation of resources—inputs—and operations to those outputs.*

In order to fully understand what this means, a distinction might be drawn. It is possible to relate all kinds of studies to the "product" or outcome or a program, or to the *operations* that attempt to produce a required result. For example, most studies still concentrate upon the inputs or operations rather than upon outcomes, partly because output-oriented studies are typically much more difficult. Hence, the administrator's report—or his budget request—will tell of the numbers of dollars, pesos, escudos, or guaranies, spent; the number of children that attended the school; the number of teachers paid; etc. All of these are accounts of resources expended in an attempt to produce educated children, in some useful sense. There will also be accomplishments in terms of process: teachers hired, teachers trained, schoolrooms built, textbooks purchased. None of these measure the changed abilities and attitudes that are the objectives of the educational system.

It is possible and profitable to study these operations in detail. A whole discipline of applied mathematics—including statistics—has been evolved to study operations in the service of management: "operations research". It has a broad compass of techniques that have not even been suggested above: scheduling, inventory control, all sorts of "optimizing" methods in the search for the best way of accomplishing an *operation*, not of producing an output. Among the useful techniques are devices, such as *critical path methods*,

which can assist in the most economical and least delayed completion of a one-time operation to be finished by a pre-assigned date. The application of methods of operations research to the management of educational systems is not too far advanced; it should be given more attention—but this is not the subject of the present seminar.

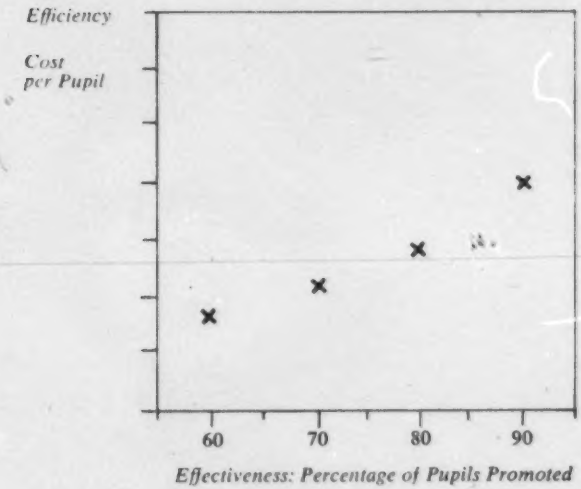
How, then, are we to apply the study of outcomes to the improved administration of educational systems?

First we must know what the outcomes of the educational system are to be, and we must measure them.

FIGURE V-7

ESTIMATED COST PER PUPIL AND PERCENTAGE OF FOURTH GRADERS PROMOTED TO THE FIFTH GRADE

(Illustrating a Likely Relation between Efficiency and Effectiveness of Educational Outcomes)



(Hypothetical Data)

The term *efficiency* is often used to indicate the cost-per-unit-of-output, such as the funds spent or the number of teacher hours per pupil who moves from the fourth grade to the fifth grade. Notice that this is different from cost accounting, which typically assigns costs to inputs, and allocates inputs to various uses; here we deal with cost related to *output*, and cost may be defined, in turn for each of the major types of resources utilized to produce the output.

Effectiveness is usually defined as a measure of the success in achieving a defined purpose (for example, the number of fourth graders that complete the work satisfactorily and proceed to the next grade). It should be apparent that efficiency and effectiveness are often conflicting objectives: it is possible to achieve cost reductions if only the easier cases (translate: children who can complete the work of the fourth grade without difficulty) are served. It is possible to draw a graph (Figure V-7) for the average cost (efficiency) of achieving different degrees of effectiveness. If the children that are difficult to teach (i.e. have learning problems) are to complete their work, they require extra effort—resources—in order to meet this objective. (See page 37).

Management-by-objectives is strictly a method of stating in advance the *output* goals to be achieved, and then exerting every effort to achieve these ends. It is also used as a device to state "process" goals (hiring 17 more teachers—a process measure—rather than retaining 87 percent of the fifth graders for the sixth grade—an outcome measure). In this ambiguity, one must first determine whether the management-by-objectives in any particular instance is in terms of outputs or whether a looser construction is being assumed.

Cost-effectiveness studies are among the principal reasons for measuring outputs in the first place. Such studies seek to find the combinations of policies and resources that produce a particular outcome at the lowest cost. Hence, in order to create a situation wherein 80 percent of the entering first grade pupils will satisfactorily complete eight grades of schooling (elementary or primary school in some countries, part of the secondary school grades in others), there are many choices of educational technology: different teaching materials, standards for teacher recruiting and training, fee schedules for the pupils, etc. There are also associated policies with any technology: free textbooks or pupil-supplied books, teacher salaries, standards for promotion from one grade to the next, etc.

What combination will produce the highest score—children retained until the completion of eight grades—per unit of resources expended?

The systematic exploration of that question demands more disciplined experimentation and assessment—and innovation—than most school systems in the world have achieved, as of this date. Perhaps one of the outcomes of this program of training by Georgetown University will be the encouragement of more purposeful experimentation in better ways of achieving meaningful outcomes of educational systems among the various countries of the world.

Benefit-cost analyses involve additional problems of great technical difficulty. In measuring cost-effectiveness, we started with an agreed-upon measure of outcome; the only question was how to achieve it at the lowest cost in resources expended, or inputs. The child with a completed eighth grade was equally valuable to society, his community, his family, and himself no matter how the result was achieved.

In benefit-cost analysis, we seek to *compare different kinds of programs or social services.* Hence, we seek to put a *value* upon different services that can be supported with the same scarce resources (funds, staff positions). How to compare three programs: one involving educational attainment, one involving improved material-and-child-health services, and one seeking economic development with its associated higher-paid jobs and enhanced occupational mobility of the population. All of these goals or objectives are worthy ones; no one in his right mind could argue against their worth. Yet it is the essence of budgetary operations to allot resources to some programs, and—in the same process—by whatever name, to exclude others or to provide lesser resources than would otherwise be available. It may be unpleasant for the ordinary citizen to contemplate, and difficult for the government official to decide, but more education (at any level of cost-effectiveness) means less maternal and child health, and less economic development. In practice, the problem is not in choosing only among the outcomes-with-values-assigned—benefits—for three alternatives, but the practical problem of choosing from among *all of the programs* or activities that seek support from the same pool of funds, personnel, and governmental authority.

It need hardly be recorded that this part of output-oriented administration is frightfully underdeveloped. What studies have

been done tend to evaluate *only in money* that can readily be related to the (unevaluated) special outcome measures of the various programs: the expected additional tax collections associated with more educated people through their lifetime; the incremental gain in Gross National Product for various differential gains in educational attainment; the avoidance of future welfare costs for better-trained workers; the savings of hospital and other medical costs for programs in preventative medicine and early treatment of potentially serious medical conditions. What is missing? Everything that cannot readily be measured by money, especially in terms of future governmental expenditures or additional governmental income thru the various tax mechanisms. This omission is very great in any terms except the fiscal measures of governments: the health program gave no credit—no value system—for the rewards of good health for the individual, and the related rewards for his family; for the richer and more satisfying life of the better educated; for the satisfactions of a highly productive profession or occupation. In short, we have neither valid systems for measuring the *quality of life*, nor even useful conventions for such measurements. Hence, we have different output measures for each kind of program, and each component of each program—such as an additional year of early grades in school and of secondary grades.

In short, the way has been prepared—if there is sufficient determination and intellectual discipline—for cost-effectiveness studies, seeking the "best" way to maximize outputs for individual programs.

The "best" answer will always remain the "best", no matter what value one puts on the outcome. It is when *different kinds of programs* are to be compared that the absence of a global value system is felt. Much of the demand for studies of benefits seem to completely ignore this problem, since there is no discussion at all of the technical issues in valuation of unlike outcomes from unlike programs. There is much literature about "bringing about change", as though the personal reluctance of the opposition of officials presented the only obstacle.

Another symptom of the problem of measuring value of outcomes is the attention given to problems that *can* be solved, such as the treatment of values for future years. If a program—say vocational education of a particular sort—can be assigned a "value" (additional wages, additional taxes associated with the incremental wages, etc.) for each year in the future, one can apply fairly standard principles of financial analysis and actuarial science to give an estimated

"present value" to each individual year, even though the gain may be many years off. Each year's benefit (however measured) is "discounted"—reduced—because future payments are deemed to be worth less than income received today. Hence we have the refinement for the part of the benefit that is converted into a common currncy of money, while ignoring what many analysts might regard as a much larger component that is too difficult to measure. It is to be hoped that more satisfactory conventions can be evolved for the valuation of various aspects of the "quality of life". Until this is accomplished, the acceptable strategy will have to be the assignment of money values for benefits wherever this can be contrived in any reasonable manner. This result will be regarded as an *understatement* of the true benefits, often with a listing of the factors (good health, rewarding occupations, a fuller intellectual and artistic life, etc.) that were left out, leaving to the reader—or the budget authorities—the job of deciding how much the scaled measure of benefits is understated.

Output-oriented management has certainly made many contributions to management and to the management of public enterprise especially. If used with some depth of understanding, and a proper sense of humility, it is bound to contribute a good deal more in the highly important business of public-administration decision making. It sometimes takes on a bit of the character of a man made monster out of control, when it is in the hands of the untrained and the simplistic. In this respect it is similar to uncounted other techniques available for management of the public business.

41

APPENDIX A

Selected Uses

As indicated in the Preface, some of the materials and "figures" in the six seminars may be particularly useful for the specific groups with whom the discussion are scheduled.

To some extent, selections will be made at the discretion of the discussion leader and may depend in part on the length of time available and on the amount of material included in the seminar.

Brief Comment on Selections from Seminars for Different Groups

Generally, Seminars III and VI, along with introductory material from Seminar I, are the most important for teachers, officials directly concerned with school administration, and teacher training institute instructors.

Seminars II, IV, and VI are most important in providing basic materials for top officials concerned with overall policy planning, middle management officials in Ministries of Education and Finance, and training institute instructors who teach planners and school administrators. Seminar I provides a backdrop discussion of concepts and of the multiplicity of outcome measurements. By and large, Seminar V will help inform top officials in planning and in Ministries of Education and Finance of what they can expect from the analyses.

Selected Uses for Seminar V

The material in Seminar V will be most useful to administrators and to people in administration supportive roles such as budgeting, planning and management. It would be effective also for use in teaching institutions preparing personnel for these functions.

This Seminar would not be useful or appropriate for teachers, teacher trainees or school principals.

APPENDIX B

FIGURE V-1

TEST SCORES OF GIRLS OF AGE 10 YEARS WHO ARE IN FOURTH GRADE (Illustrating a Distribution)

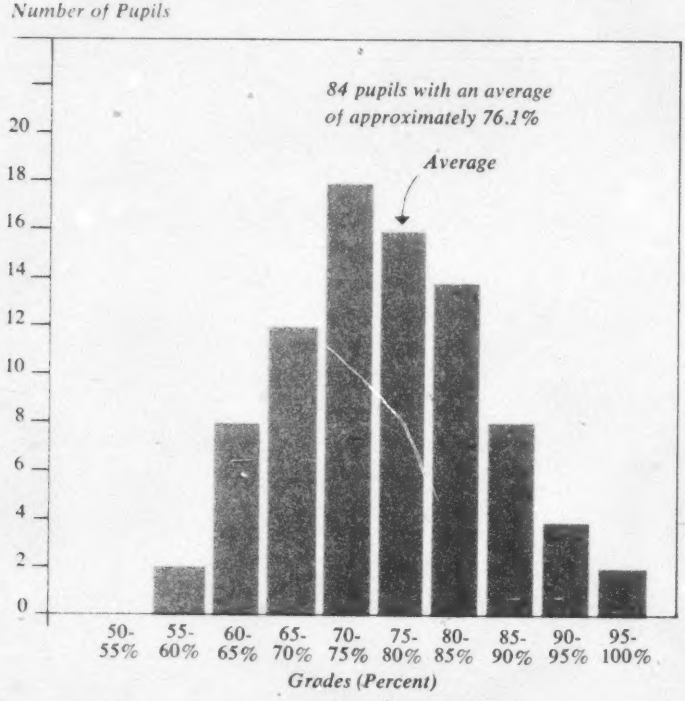
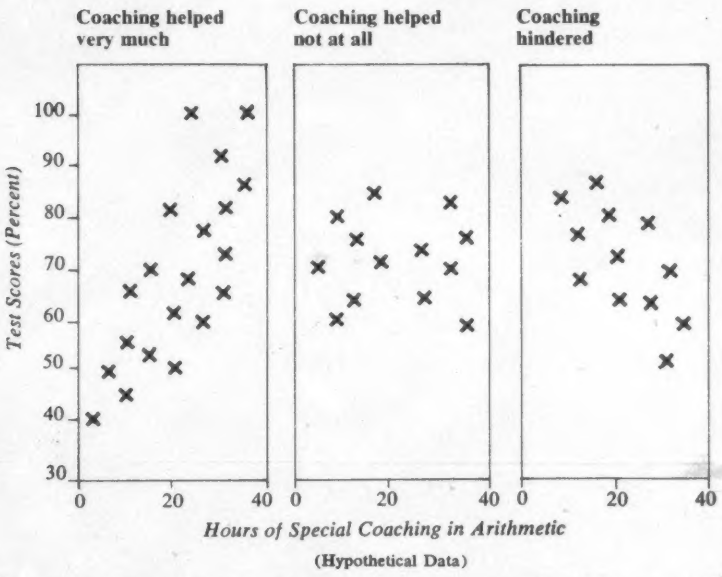


FIGURE V-2

RELATION OF ARITHMETIC SCORES TO HOURS OF COACHING (Examples of Scatter Diagrams)

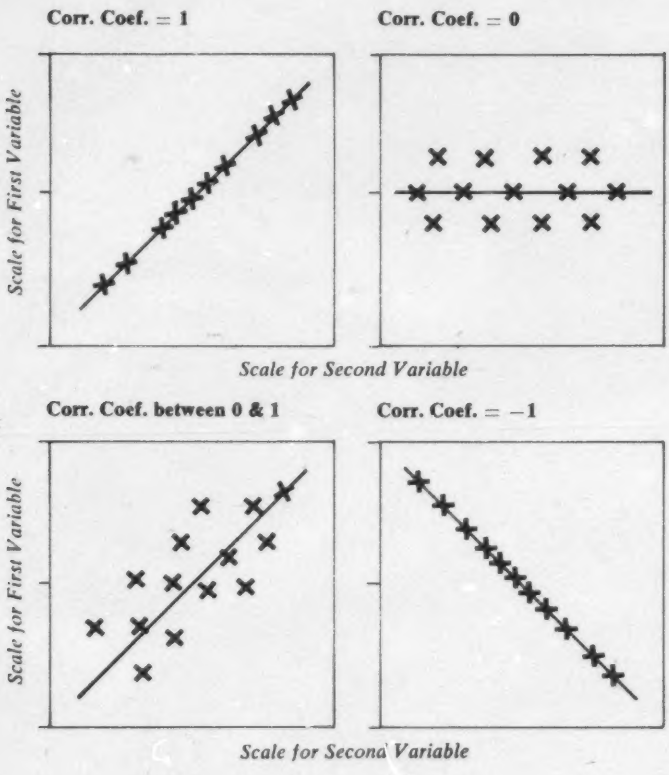


(Hypothetical Data)

FIGURE V-3

CORRELATION BETWEEN TWO VARIABLES

(Illustrating the Correlation Coefficient)

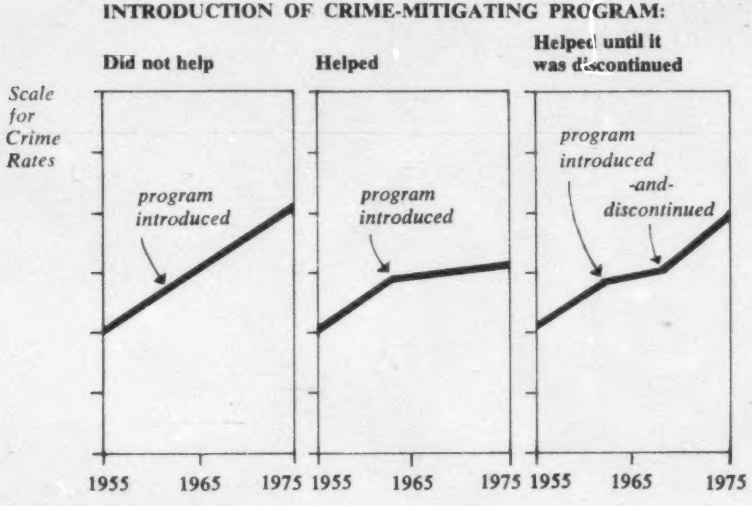

Corr. Coef. = 1
Corr. Coef. = 0
Scale for First Variable
Scale for Second Variable
Corr. Coef. between 0 & 1
Corr. Coef. = −1
Scale for First Variable
Scale for Second Variable

(Hypothetical Data)

FIGURE V-4

EFFECTS OF CRIME-PREVENTION PROGRAMS

(Illustrating Trend Analysis)

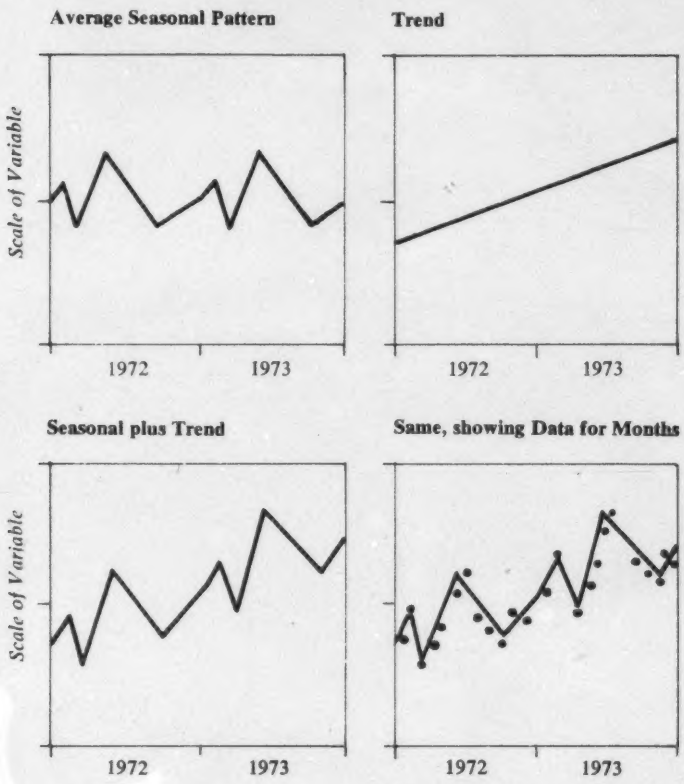


(Hypothetical Data)

FIGURE V-5

VARIATIONS OVER TIME

(Illustrating Components of a Time Series)



(Hypothetical Data)

FIGURE V-6

TEST SCORES OF CHILDREN FROM SCHOOLS IN TWO COMMUNITIES

(Illustrating "Confounding" of Explanatory Influences and Sensitivity of Statistical Analysis)

Children whose Parental Educational Attainment is:	*Schools from Communities where Average Family Income is:* **High**	**Low**
High	214: 83%	2: *
Low	4: *	307: 72%
High	50: 83%	200: 81%
Low	200: 75%	50: 72%

Key: Each cell shows number of children taking the test and their average score

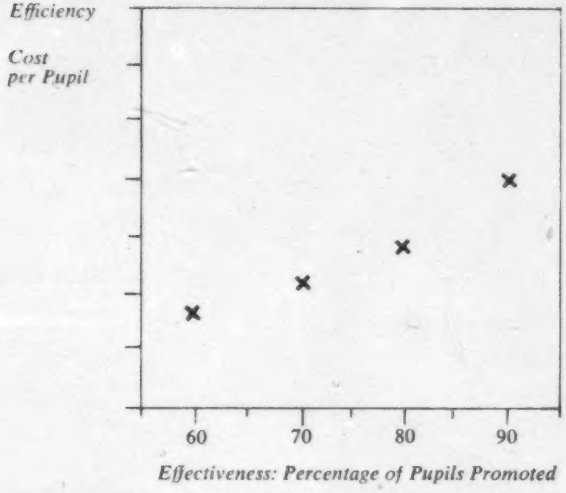
(Hypothetical Data)

**Sample too small for average to be significant*

FIGURE V-7

ESTIMATED COST PER PUPIL AND PERCENTAGE OF FOURTH GRADERS PROMOTED TO THE FIFTH GRADE

(Illustrating a Likely Relation between Efficiency and Effectiveness of Educational Outcomes)



(Hypothetical Data)